Towards Owners' Control in Digital Data Markets

Sabrina De Capitani di Vimercati, *Senior Member, IEEE,* Sara Foresti, *Senior Member, IEEE,* Giovanni Livraga, *Member, IEEE,* and Pierangela Samarati, *Fellow, IEEE.*

***Abstract*—Finding a good balance between the availability of data for analysis and the control that individuals should exercise over their own data is a key requirement for generating innovation and growth in our data-driven society. A promising direction is the development of data market platforms where individuals can directly provide their own data and monetize them by making them selectively available to data consumers. Data market platforms are often based on the cloud paradigm and can be managed by parties that may be not fully trusted, or even be malicious, thus introducing new data security and privacy issues. In this paper, we discuss the issues and challenges towards empowering individuals to use a data market platform for trading their data while keeping control over them. We also discuss how existing techniques can be possibly adapted to address these issues and highlight aspects that still need to be investigated.**

***Index Terms*—Data market, Data protection, Owner control, Privacy, Security**

I. Introduction

It is a well-recognized fact that our society and our economy depend more and more on data. The numbers behind the so-called data economy are astonishing and grow at tremendous rates. In the EU alone, the value of data economy is expected to reach €739 billion by 2020, representing the 4% of the overall EU GDP (from €285 billion in 2015, around the 1.94% of the EU GDP) [1], and ensuring free flow of data is seen by many as a fundamental driver for innovation and growth. When it comes to personal (or sensitive) data generated by citizens, the scenario becomes more complicated. On the one hand, personal/sensitive data make no exception in having a value that can be cashed. On the other hand, recent laws and regulations, such as the General Data Protection Regulation (GDPR, https://eur-lex.europa.eu/eli/reg/2016/679/oj) and the forthcoming ePrivacy regulation in the EU, clearly aim at giving control to data subjects (i.e., individuals to whom data refer) over their personal data. In principle, one may argue that personal data should not be used for generating profit, and hence the data economy should only concentrate on public data. Still, personal data are constantly gathered and analyzed, as testified, for example, by scandals related to the collection and analysis of personal data to profile and psychographically segment unaware individuals [2]. This has been showed to permit to deliver more effective advertisements or political messages thus generating profit, which most of the times benefits a few big players on the market (e.g., the providers of social media or online services) to which individuals give access to their personal data in exchange for their (so-called) 'free' services. Hence, the increasing call for practical and efficient solutions that empower individuals with control over the sharing of their own personal data [3].

Towards this goal, a relatively new idea that is gaining popularity is that of a paradigm shift from the current scenario where data of individuals are owned, controlled, and monetized by companies, to a scenario where data are controlled and monetized by their owners. A possible solution concerns the development of *data market* platforms, where individuals can directly provide their own data, and decide on whether, how, and for what purpose selectively share them with interested third parties and receive a reward for that. The benefits of such an approach are multiple and can have a positive impact on both the individuals and the third parties purchasing data. On the one hand, keeping control over their own data allows individuals to have a share on the money that the usage of those data can generate while maintaining privacy. On the other hand, the use of a data market platform allows third parties to obtain access to personal data in accordance to the wishes of the individuals and in a transparent way, thus boosting a circle of trust that could, in turn, encourage more individuals to contribute their data to the market. Also, having a centralized market platform can create opportunities for the individuals, who can enjoy a window for publicizing their data to interested entities, and third parties, who can easily find, in a large pool of data, the ones that are more of interest to them. The possibilities of such data market platform are countless for the individuals who can then, depending on their wishes, decide to use it for trading raw data (e.g., data generated by IoT/wearable/smart devices), sanitized/encrypted data over which some protection mechanisms have already been enforced (e.g., private statistics or aggregates computed locally) as well as data resulting from analyses or computations, performed by the market provider (if considered trusted) or by the individuals themselves, over raw data.

The realization of such a market platform entails a number of issues that need careful analysis and proper solutions. While some of them are similar to the problems that can arise in a scenario where datasets are outsourced to external (e.g., cloud-based) platforms, and mostly derive from the loss of *direct* control over data, others are specifically related to the peculiarities of the data market scenario. The goal of this paper is to analyze the issues and challenges that characterize the development of a data market platform that allows individuals and interested third parties to interact for trading personal data, while ensuring that individuals remain in control of their own data. In the remainder of this paper, we first illustrate and characterize the reference scenario, with its subjects and the different trust assumptions that can hold on them (Section II). We then discuss the issues and challenges entailed by the considered scenario (Section III). For each issue, we provide

Sabrina De Capitani di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati are with the Computer Science Department, Università degli Studi di Milano, Italy. E-mail: *firstname.lastname*@unimi.it

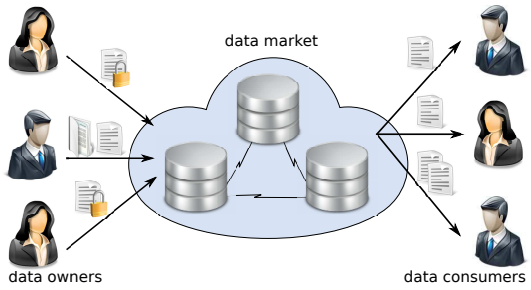

Fig. 1. Reference scenario

a brief description of existing techniques along with open challenges that still need to be addressed.

II. Scenario and Trust Assumptions

The reference scenario, graphically illustrated in Figure 1, is characterized by multiple parties interacting on a (public) platform, the *data market*, to the aim of trading data for something of value (e.g., money). The data market often relies on the cloud paradigm for providing data storage and computation services and is managed by a *market provider* in charge of its maintenance. The data market offers a number of services to its customers, who can play as *data owners* and/or as *data consumers*. Data owners are interested in monetizing their data on an open market. Note that we refer to data owners for generality: a data owner can be any entity that can have the right to share a data item (e.g., private individuals and private companies). Consumers are interested in purchasing datasets and/or analysis or computations over them through a data market (e.g., producers interested in consumers habits, fitness centers interested in sport preferences). The data market provides access to a wide availability of data from several sources and enables data exchange between owners and consumers, providing the infrastructure needed for data storage, transfer, processing, and for payment management.

An important aspect that influences the security issues to be addressed as well as the techniques needed for properly protecting data is related to the *trust assumptions* on the parties involved in the data market scenario, as they could cheat or collude to try to increase their profit. Both data owners and consumers can be fully trusted or malicious. Fully trusted data owners aim at monetizing their data and are therefore expected to be willing to provide data of good quality, to the benefits of their reputation. Fully trusted consumers aim at searching for data of interest and using them according to possible restrictions imposed by the data owners and/or contracts stipulated, for example, with the market provider. Malicious data owners could aim at performing malicious manipulation of data by providing fake data or data of low quality. They aim at having an immediate economic advantage from selling these data or at introducing biases in the analysis over certain kinds of data (e.g., market analyses). Malicious consumers can use data beyond legitimate purposes, abusing them. For instance, this could be the case of a consumer declaring the need for accessing data for research purpose, while actually using them for marketing. We assume that data owners and consumers interact through the market platform. Since the market provider is an external third party, which is under the control of neither the data owners nor the consumers, the possible trust assumptions related to the provider are similar to the assumptions usually considered to outsourcing scenarios [4]. A data owner (and/or consumer) can then consider the market provider to be characterized by different levels of trust as follows.

- *Fully Trusted*: the provider is fully trusted to behave correctly, ensuring data availability as well as data and computation integrity, and it is also trusted to access the content of the owners' data and to observe actions performed by different parties over them.
- *Honest-but-Curious*: the provider is considered trusted to correctly manage outsourced data (i.e., it is trustworthy with respect to integrity and availability), but it is not trusted to access the (possibly sensitive) content of the data, as well as to observe actions over them performed by owners and/or consumers, since they might reveal sensitive information.
- *Lazy*: the provider is not considered fully trustworthy with respect to integrity and availability since it could aim at saving storage and/or computation resources (e.g., the provider could return the result of a computation without actually performing it).
- *Malicious*: the provider is not considered fully trustworthy with respect to integrity and availability as it might act maliciously when managing and processing data (e.g., the provider may return an incorrect computation result with malicious intent or may be a victim of attacks guided by an external party aimed at sabotaging the data market).

Data owners, therefore, need techniques that can effectively protect their data, especially in situations where the market provider is honest-but-curious, lazy, or malicious. Data protection implies guaranteeing data confidentiality, availability, and integrity, the latter being a crucial issue when the market provider could possibly modify the owner's data. Analogously, consumers might need to protect the confidentiality and integrity of their actions over the data with respect to the market provider (as well as any other observer). We note that the trust assumptions on the market provider can be asymmetric: different entities (be them consumers or data owners) might trust differently the same provider.

III. Issues and Challenges

The peculiarities of the scenario described in Section II raise several issues and challenges that need to be addressed to enable both data owners and consumers to take full advantage of data market platforms. These challenges are related to the definition of a market platform able to easily support the storage, processing, and exchange of (possibly sensitive) data and to manage economic transactions due to data trading. The efficiency of the data market platform is a cross-cutting problem that affects all solutions developed in this scenario. As a matter of fact, the platform should have virtually no downtime and should guarantee fast and responsive interactions, so that data can be uploaded, published, downloaded, and accessed

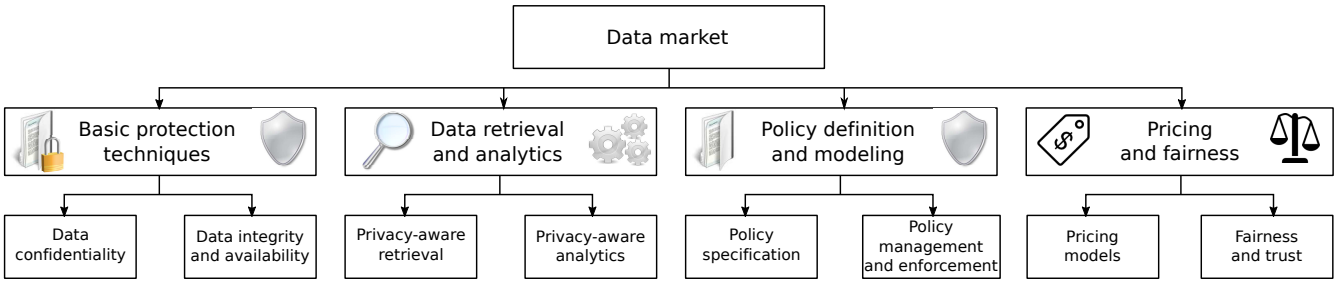

Fig. 2. Taxonomy of the challenges discussed

with no delays, whenever needed [5], [6]. Such interactions could involve a large number of data owners and consumers and should be characterized by low latency. The success of a data market platform is also determined by its scalability, as the needs of all data owners and consumers should be met.

In the remainder of this section, we give an overview of these main issues and challenges, which are summarized in Figure 2. First, we focus on the need for basic protection techniques, aimed at guaranteeing confidentiality, integrity, and availability of the data traded in the market (Section III-A). Second, we analyze the need for solutions enabling data retrieval and analytics, while ensuring privacy of sensitive information stored in the traded data and of users' actions (Section III-B). Third, we consider the need for specifying and enforcing user-based policies regulating access to the data (Section III-C). Finally, we discuss pricing and fairness issues that characterize markets where products or services are traded in exchange of money (Section III-D).

A. Basic Protection Techniques

An important need of data owners wishing to monetize their data is the availability of a market platform for storing, managing, sharing, and analyzing such data. However, delegating the management of data to the market platform introduces the issue of properly protecting data against external parties (possibly including the market provider). Data protection requires the design of basic protection techniques aimed at guaranteeing data *confidentiality*, *integrity*, and *availability* (Figure 2).

Data Confidentiality. Confidentiality requires that sensitive data be protected against unintended disclosure, also to the market provider if not fully trusted [7], [8]. The confidentiality problem in the data market scenario resembles the analogous problem in outsourcing and cloud scenarios where data are typically managed by an external party. For the data market scenario, there is then the possibility to build on the techniques proposed in these scenarios, possibly adapted as needed. In particular, for the data market scenario two complementary approaches can be adopted, depending on the kind of transformation performed over the data. On the one hand, *encryption-* and/or *fragmentation-based approaches* aim at protecting data by (partially or completely) disabling visibility over data while preserving functionality that can be enjoyed by both data owners and consumers. On the other hand, *data sanitization approaches* (e.g., k-anonymity [9], differential privacy [10]) aim at providing an obfuscated version of the data or non-sensitive aggregates over them.

Encryption-based approaches typically rely on owner-side encryption to maintain data private to the storing provider. Since encryption affects query execution, researchers have investigated techniques that support the execution of queries directly on the encrypted data. Existing solutions rely on *homomorphic encryption*, which allows the execution of a computation directly over encrypted data (e.g., [11]), or special *property-revealing encryption* schemes (e.g., searchable symmetric encryption, deterministic encryption, order-preserving encryption), which support some operations (e.g., [12]). Alternative solutions are based on *indexes* (e.g., [13]), or on the use of *encrypted database systems* (e.g., [14]). Indexes are metadata associated with the encrypted data values that are used by the provider to select the data that satisfy a query condition. Indexes can be based on, for example, hashing, deterministic encryption, tree-based structures (e.g., B+-tree). Different kinds of indexes have been investigated, including: *direct indexing*, which provides a one-to-one correspondence between plaintext and index values; *bucket-* or *hash-based indexing*, which provides a many-to-one correspondence between plaintext and index values; and *flat indexing*, which provides a one-to-many correspondence between plaintext and index values. Different kinds of indexes ensure different protection guarantees as well as different support for query execution. For instance, since bucket- and hash-based indexing techniques provide a many-to-one correspondence, meaning that different plaintext values collide to the same index value, the result computed over such indexes may include spurious data due to collisions that must be filtered out by the query submitter. Encrypted database systems support the execution of SQL queries directly on encrypted data. An example is CryptDB [14], a database system based on a multi-layer encryption technique. Each column of a relational table is protected through different layers of encryption, where each layer supports different kinds of SQL queries. Whenever the system receives an SQL query, it determines the layer of encryption supporting the query and, if the encrypted data are not already at the needed layer, the system removes the other layers until the one needed is reached. The system then executes the query and returns the result.

Fragmentation-based techniques have been developed to reduce the use of encryption (e.g., [15]). The idea is that, whenever what is sensitive is the association among values

rather than values singularly taken, data can be partitioned in different fragments in such a way that each fragment does not include sensitive associations. As an example, suppose that a data owner is willing to sell her daily activity data collected by her fitness tracker, which include, among others, the heart rate and footstep patterns. Suppose also that the data owner considers the association between heart rate and footstep sensitive. Instead of encrypting the whole dataset, the data owner could decide to split heart rate and footstep in two fragments that cannot be joined.

Sanitization techniques produce an obfuscated version of a dataset. These approaches can be categorized in two main families: *syntactic approaches*, which build on k-anonymity [9] and its variations, and *semantic approaches*, which build on differential privacy [10]. Syntactic approaches provide protection by decreasing the level of detail in a dataset, to hide the data of an individual (be them her identity or sensitive/personal information) in a group of individuals in such a way that an observer (i.e., the market provider or a consumer) cannot associate a data item with a single individual. Semantic approaches perturb the result of a computation (e.g., adding noise) to ensure that the analysis over the dataset returns similar results independently from the presence or absence of the data of a certain individual in the dataset used for the analysis.

While promising, the solutions discussed above still need to be refined for the data marker scenario. For instance, indexes and encryption supporting queries may cause information leakage [16] and provide limited support for the execution of generic computations. Traditional sanitization approaches operate on a dataset in its entirety, so they are applicable when the entity obfuscating it (e.g., the market provider) is trusted to access the original (unprotected) data. A possible direction in this regard concerns the consideration of *local differential privacy* [17], where noise can be applied *locally* by each data owner to her own data collection before sending it to the market provider. Furthermore, the protection mechanism must be efficiently enforced and must be compliant with the possible requirements defined by the data owner. Such requirements may impose constraints on how data can be searched, accessed, shared, and processed (e.g., a given dataset can only be shared in aggregated/sanitized form [18], or only for specific purposes – see Section III-C).

Data Integrity and Availability. Integrity requires the authenticity of the data stored in the market as well as of the manipulations over them. For instance, modifications and/or deletions by the market provider or by unauthorized parties should be detected and possibly prevented [4]. Availability requires that data be readily available when needed and be managed according to possible requirements imposed by the data owners. For instance, if a data owner wishes to sell her data to non-profit organizations, such organizations should be able to access the data without undue delay or limitations.

With respect to integrity, some approaches have investigated the problem of providing guarantees that data are correctly stored at external providers. Traditional solutions are based on the use of digital signatures, meaning that each piece of data is signed by the data owner with her private key. To make the signature verification process efficient, multiple digital signatures can be combined according to different approaches (e.g., condensed RSA, batch DSA signature aggregation, or BGLS [19]). Provable Data Possession [20] and Proof Of Retrievability [21] allow a provider to produce a proof that a verifier (in our context, data owner or consumer) can use to check if data are intact (integrity) and if the provider still stores the data previously uploaded (availability). With respect to availability, existing solutions are typically based on replication, meaning that data are distributed redundantly across different providers [22], and on the definition of contracts (Service Level Agreements) where the data owner and the market provider agree on the characteristics contributing to the availability of the platform and of the data (e.g., number of replicas).

The solutions discussed above can also be applied in the data market scenario. However, the choice of the specific solution to be considered may depend on the privacy needs of the data owners as well as on the trust assumption on the market provider. For instance, if a data owner does not want to reveal her identity in association with a dataset, signature-based solutions cannot be adopted.

B. Data Retrieval and Analytics

Moving data to a market platform enables the creation of a centralized data hub (which however does not imply physical centralization) that can facilitate data visibility for interested consumers. Indeed, consumers would have an easy access to a data market where they can find the data and/or analytics of interest. This also naturally reduces the burden left to the data owner who does not need to be always available for responding to consumers' requests. Like for the storage of data, the retrieval and the computation of analytics imply the need of addressing specific privacy issues (Figure 2).

Privacy-Aware Retrieval. Data market platforms should be easy to use by both data owners and consumers and should facilitate data trade between them. This implies that the platform should have a *market catalog dashboard* that consumers can use to quickly browse the datasets available and easily match their demand with data owners' offers. In particular, the market catalog dashboard should keep track of all the data stored in the market in association with searchable characteristics (metadata) such as price, keywords describing the dataset, creation date. This problem has some similarities with the problem of exposing web services via a catalog in the Service-Oriented Architecture paradigm to support users in choosing (and possibly composing) the best services accommodating their needs (e.g., [23]). A complicating factor arising in the data market context is that such a catalog should take into consideration the need for confidentiality: the datasets available on a market should be properly included in the catalog, while not exposing data content to unauthorized consumers. Open challenges involve the definition of privacy-preserving metadata associated with the datasets, and the adoption of efficient privacy-preserving search techniques operating on them. In fact, a consumer may be interested in protecting her searches since their target may reveal something about

her intention. As an example, suppose that, before launching a new wearable device on the market, a producer wants to buy and analyze the datasets collected by competitors. By observing the access requests posed by the producer (acting in this case as the consumer), an observer (including possibly the market provider) may infer that the producer may be willing to develop a new device. Existing solutions for protecting the privacy of access patterns are based on adaptations of the Oblivious RAM data structure (e.g., [24], [25]) or on tree-based data structures such as the B+-tree in the shuffle index approach (e.g., [26], [27]). While interesting, these techniques cannot be directly applied in the data market scenario, since they work under the assumption that the querier has direct control over the whole data collection, which is however not the case of a consumer in the data market.

Privacy-Aware Analytics. The power of data lays in the information and knowledge that can be derived from analyses over them. Besides being used for trading raw data, data markets should also have functionality for the computation of analytics over the data, possibly coming from multiple sources and/or owners. As a matter of fact, data of different owners may have an enormous value whenever they are combined, aggregated, and processed together. For instance, consider the data collected by wearable devices, which are typically related to the habits of their owners. In this case, if the data of different owners are analyzed, each owner will be able to compare her performance, for example, with the performance of other people of the same age. This has advantages not only for the product owner but also for product manufacturers, who can develop prediction models able to better support their customers. With respect to the kind of analysis that can be supported on the data, we envision two different scenarios: *i)* the market provider can perform pre-defined analyses, whose results can then become other resources sold on the market; and *ii)* the market provider can respond to arbitrary requests coming from interested consumers, who should then be able to specify the kind of analytics in which they are interested, and on which data the analytics should be performed. Both scenarios are complicated by the fact that the data owner may want to impose some restrictions on the kind of computations that can be executed or on the kind of protection to be enforced on her data (Section III-C). For instance, a data owner might specify that her data could be used only in analyses whose results correspond to aggregates computed on at least 1000 distinct records. Although several approaches have been designed to protect the privacy of the users (e.g., [18]), this problem is particularly difficult in the data market scenario because the analysis may involve data of different owners with different (possibly contrasting) requirements that should be enforced. Also, since the result of an analysis over a (collection of) dataset(s) can leak sensitive information on the data that have been analyzed (and hence on the individuals to whom data refer), the data market should support data trading while protecting the privacy of all involved parties, thus avoiding both direct and indirect leakage of sensitive information (e.g., [28]).

In addition to the confidentiality/privacy problem, also the integrity of the analyses is an interesting aspect that should be considered. In fact, a lazy or malicious market provider may be interested in reducing the operation cost to maximize its profit. The data market provider could then operate on a subset of the raw data, or produce a fake result without working on the raw data. The integrity of computation results is a well-known problem and existing proposals aim at ensuring the *completeness*, *correctness*, and *freshness* of the result of a computation. Completeness means that the computation has been performed on the whole dataset and the result includes all the data of interest. Correctness means that the computation has been performed on the original dataset and the result is correct. Freshness means that the computation has been performed on the most recent version of the dataset. Current solutions are based on the definition of *authenticated data structures* (ADSs) or on *probabilistic controls* [4]. ADSs are data structures (e.g., skip lists, Merkle trees and their variations) built over an attribute of the dataset that can support the verification of the integrity of equality and range queries executed over the attribute on which the structures are built. Probabilistic techniques are instead more general since they can support the verification of the integrity of any computation but provide only a probabilistic guarantee. These latter techniques rely on the controlled injection in the dataset of ad-hoc control information (e.g., fake or replicated data), whose absence (or incorrect evaluation) in the computed result signals an integrity violation. While interesting, these techniques are not directly applicable in the data market context since they can expose the identity of the data owner (which may need to be protected) and may not support generic computations/aggregations. Also, integrity guarantees may be more difficult to verify when the computation is performed by the market provider through the combination of datasets of different data owners.

C. Policy Definition and Modeling

Data sharing and dissemination are basic features for data markets that should be supported in a controlled way so that data owners remain in control of their data. This is also in line with recent laws and regulations (e.g., the EU GDPR) that empower the subject of a data item (i.e., the individual to whom the data item refers) with rights over it. For instance, the GDPR states that a certain data item should only be used for purposes for which its subject agreed, and that it should be deleted without undue delay should the subject request so. The consideration of these problems in the data market scenario introduces the need for supporting expressive and flexible user-based policies, which impose restrictions on the use and processing of data, and efficient and effective enforcement mechanisms (Figure 2).

Policy Specification. The need for empowering data owners to be informed on the use of their data and be active participants in this process is largely recognized. The growing interest in data market platforms for trading personal data clearly strengthens such need, calling for solutions that can regulate the use, processing, and dissemination of personal data, enhancing the control of data owners on their data. Recognizing

the importance and interest of these techniques, the research and industrial communities have investigated solutions for empowering users with control over their information in digital interactions and, in general, in data collection and processing (e.g., [29]). Most proposals have focused on solutions that the party offering a service/resource (provider) can use, and assumed to adopt similar approaches at the data owner side. However, these approaches do not work well because data owners may need a way to specify preferences/restrictions on the use of their data. As an example, consider a data owner willing to sell the data collected by her wearable device. The data owner could outsource her data to a data market platform together with her privacy preferences saying that the data should be sold only to companies that aim to carry out research focused on the detection of "human features" such as sleep and stress. This problem could be addressed by defining a simple, flexible, and expressive model and language (suitable for end users) supporting the definition of regulations/preferences by data owners on their data. A similar problem has been traditionally considered in the context of open systems. In fact, some proposals have focused on the user perspective and allow users to selectively release their attributes/properties in the interaction with the system (e.g., [30]). Some attention has been also devoted to the problem of empowering the user to specify privacy restrictions, regulating the secondary use or the dissemination of the data released by the user to satisfy an authorization policy and acquire access to services or resources (e.g., P3P), as well as data handling restrictions, which are attached to the data, regulating the use of users' data by the receiving parties (e.g., [31]). While interesting, all these proposals consider only some aspects of the problem and cannot easily fit the peculiarities of the data market scenario.

In the data market scenario, a possible solution for addressing the problems above-mentioned consists of using contract agreements and obligations. However, contracts and obligations may only partially empower the data owner, since they do not provide any guarantee on the proper protection of data and on the ownership control. A full realization of the data market scenario requires to define solutions that allow data owners to specify policies regulating protection of data and their selective disclosure in the market itself. More concretely, there is the need to provide a usable language for expressing advanced protection needs such as: sharing options, indicating the parties with which the data can be shared; the granularity with which specific information can be analyzed, processed, and shared; context-based restrictions, or actions that should be triggered when simple context-based conditions are satisfied; the purpose of the request; and the history of accesses already purchased (which, in turn, implies the capability of reasoning on inferences that all the accesses bought in a certain time window can open). The model and language should also support the specification of different kinds of requirements (e.g., secondary use, provenance, legal requirements [32]) that can be defined on both the raw data and the result of an analysis (e.g., allowing a purchase only for non-profit research). Such requirements can also be related to the specification of restrictions forcing the use of sanitized or obfuscated (e.g., encrypted) versions of data for release to third parties or usage in an analysis, so to enrich the opportunity of data sharing in full respect of the specified policies.

The definition of such a model and language will allow data owners to be more conscious of how their data are used in practice while at the same time allowing them to take an active role in their management, thus ensuring a fine-grained control on their personal information.

Policy Management and Enforcement. To ensure that data owners' policies be taken into account, there is the need of providing a means to effectively enforce the policies. The enforcement should not rely (solely) on the market provider, especially when it is not fully trusted. Recent developments in trusted hardware can be exploited to have a trustworthy component dedicated to the enforcement of policies tied to data, also in presence of lazy or malicious market providers [33].

In general, the techniques for the enforcement of policies may depend on the kinds of restrictions specified in the policy itself. For instance, as previously discussed, a policy may impose the protection of data to which it is tied through a specific protection technique. Such a technique could be applied at any stage of the data life-cycle: *acquisition*, when data are transferred from the data owners to the data market; *storage*, when data are stored and managed by the data market provider; and *analytics*, when data are processed in the data market. The use of such techniques should guarantee an acceptable degree of *utility* to the consumers and should be executed in an efficient and scalable way by the party responsible for its enforcement.

Access restrictions could be enforced by extending traditional authorization models (e.g., [34]), possibly operating on credentials in open scenarios (e.g., [35]), with additional constraints on the purpose of access and considering the peculiarities deriving from purchasing/trading (e.g., [36]). To avoid relying on the provider for authorization enforcement, existing solutions are based on the concept of *self-enforcing* access restrictions, through *selective owner-side encryption* [37]. Different data items are encrypted with different encryption keys by the data owner, and each consumer is provided with the keys of all and only the data items for which she is authorized. To limit the burden of key management, it is possible to adopt *key derivation* techniques, by means of which each authorized subject is communicated a single key from which she can derive the keys she is entitled to know (e.g., [38]). An alternative strategy to selective encryption relies on the adoption of *attribute-based* encryption, a public-key encryption scheme. A subject can obtain the encryption key for a data item if and only if her attributes (i.e., information that characterizes her such as date of birth and home address) satisfy the policy regulating access to the data item [39].

Since the data market scenario is characterized by the presence of data of different owners, a further challenge is the design of mechanisms for supporting policy matching and composition, and for orchestrating policies associated with different datasets possibly involved all together in a computation. Special attention must be posed on problems related to the policy compliance with the specified data protection, usage, and sharing requirements associated with the raw data.

D. Pricing and Fairness

Data owners should receive an appropriate reward for contributing their data to the market, both when their raw data are directly purchased by interested consumers, and when they contribute their data to the computation of analytics. Such rewards may be, for example, proportional to the amount of privacy lost when more or less precise personal data are sold to consumers. Also, access should be fair and ensure that all parties involved in an economic transaction pay/receive money according to what has been agreed. Providing an adequate pricing model and guaranteeing fairness in trading data are crucial for the success of data market platforms (Figure 2).

Pricing Models. Having methodologies for assessing the value of a dataset or of the result of a computation is a critical aspect in the context of data markets. Currently, datasets may be accessed as a whole or thorough APIs for answering queries, and existing data markets adopt different pricing schemes such as monthly subscriptions or a fee that depends on the number of transactions involved in a query. For instance, Azure Marketplace DataMarket (datamarket.azure.com) is based on a monthly subscription that can be of two types: unlimited, meaning that a consumer is charged monthly and the number of transactions on datasets is unlimited; limited, implying each month a fixed number of transactions (where a transaction can include up to 100 retrieved records). The definition of a standard model for pricing is a key challenge that would help both data owners to price optimally their data, and consumers to make comparisons to pay a fair price [40].

Ideally, a pricing model should take into consideration different parameters such as the age of the data, the credibility, accuracy and quality of the data, whether the data are provided on an exclusive basis rather than to more consumers, and so on. Although some researchers have started to address these issues, the consideration of security and privacy aspects in the definition of a pricing model is still an open challenge. For instance, an intuitive pricing model can be based on the assumption that all data items (e.g., tuples) in a dataset have the same value. However, in general, the price of a dataset might vary depending on different security and privacy aspects and/or the parties involved. For instance, a (possibly anonymized) medical dataset might be sold for a lower price to a non-profit research institution than to a private pharmaceutical company. Open challenges include the definition of pricing models obeying owners' wishes and demands, but at the same time taking into consideration different factors such as consumers' identities and/or roles, the purpose of the purchase, the history of past purchases by a same consumer, and the context in which the purchase is being performed. Also, since monetization comes at the price of privacy loss (the owner shares a data item with others, losing control over it), pricing should take into account the quantification of the privacy loss deriving from selling a certain data item or from participating in certain analytics. The same data could then be sold at different prices depending on whether/which data protection techniques are applied to the data before being sold/processed. For instance, an anonymized version of a dataset could be sold for a lower price than its original version.

An alternative for the definition of pricing models is based on auctions: data owners do not set a fixed reward for their data but follow the market demands (e.g., [41]).

Fairness and Trust. The interactions between data owners and data consumers, possibly mediated by the market provider, should be fair, meaning that the correct reward (e.g., a payment for an agreed amount of money) should be safely paid to the owners by the consumers, in exchange for accessing data and/or analytics over data. Note that trading money always requires trust among the parties. Indeed, if the payment is performed after a data owner has given access to a resource, then the data owner needs to trust the consumer to complete the payment. Conversely, if the payment is performed before the data owner has given access to the resource, then the consumer needs to trust the data owner to give access to the resource for which she paid. Since data owners and consumers might not completely trust each other, there is the need for solutions that allow them to conclude an economic transaction (i.e., the trading of a set of resources) and prevent and/or expose possible misbehaviors by any of the interacting parties.

In principle, the market provider could be trusted for mediating money exchange. However, in some scenarios, we expect the provider not to be trusted (by the data owner, the consumer, or both) for money transfer (e.g., a malicious provider might steal a payment by not routing it to the correct owner). In these cases, there is a need for automatic and secure methods for transferring money. Recent proposals for the automatic and secure transfer of money in scenarios characterized by little mutual trust among interacting subjects are based on distributed ledgers such as blockchain, and on smart contracts built on top of them (e.g., [3], [42], [43], [44], [45]). While such an approach is certainly promising in data market scenarios (where an owner could sell data to a consumer through a smart contract), there are a number of issues that need to be solved. For instance, smart contracts and their execution lack confidentiality and privacy, as plain visibility over the content of a contract and over the data it manipulates is necessary for the validation during the consensus approach (e.g., [42]). Hence, smart contracts cannot include the (plaintext) values of the traded data or analytics, since any blockchain user reading the smart contract would have access to the included data. Trading encrypted data can be an option, but the issues here come in terms of key exchange: encryption keys cannot be traded with smart contracts, and there is then the need of other secure solutions for exchanging encryption keys while ensuring that the process guarantees the delivery of the reward to the owner. Other challenges derive from the (well-known) scalability issues coming with blockchain.

IV. Conclusions

We have discussed the main issues and research challenges at the basis of the development of effective and efficient digital data market platforms for trading personal data. In particular, we have focused on the issues related to the proper protection of data and computations to concretely enable data owners and consumers to take full advantage of such platforms, while ensuring that data owners be always in control of their data.

ACKNOWLEDGMENT

This work was supported in part by the EC within the H2020 Program under grant agreement 825333 (MOSAICrOWN) and by the Italian Ministry of Research within the PRIN 2017 project 2017MMJJRE (HOPE).

REFERENCES

[1] European Commission, "Final results of the European Data Market study," https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy, 2017.

[2] C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach," The Guardian, https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, 2018.

[3] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Empowering owners with control in digital data markets," in *Proc. of IEEE CLOUD*, Milan, Italy, July 2019.

[4] P. Samarati and S. De Capitani di Vimercati, "Cloud security: Issues and concerns," in *Encyclopedia on Cloud Computing*, S. Murugesan and I. Bojanova, Eds. Wiley, 2016.

[5] R. Jhawar, V. Piuri, and M. Santambrogio, "Fault tolerance management in cloud computing: a system-level perspective," *IEEE Systems Journal*, vol. 7, no. 2, pp. 288–297, June 2013.

[6] S. Das, M. Khatua, S. Misra, and M. S. Obaidat, "Quality-assured secured load sharing in mobile cloud networking environment," *IEEE TCC*, vol. 7, no. 1, pp. 102–115, January 2019.

[7] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Practical techniques building on encryption for protecting and managing data in the cloud," in *The New Codebreakers: Essays Dedicated to David Kahn on the Occasion of His 85th Birthday*, P. Ryan, D. Naccache, and J.-J. Quisquater, Eds. Springer, 2016.

[8] G. Livraga, *Protecting Privacy in Data Release*. Springer, 2015.

[9] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE TKDE*, vol. 13, no. 6, pp. 1010–1027, November 2001.

[10] C. Dwork, "Differential privacy," in *Proc. of ICALP*, Venice, Italy, July 2006.

[11] J. Feng, L. Yang, Q. Zhu, and K. Choo, "Privacy-preserving tensor decomposition over encrypted data in a federated cloud environment," *IEEE TDSC*, 2018, pre-print.

[12] J. Feng, L. Yang, G. Dai, W. Wang, and D. Zou, "A secure high-order Lanczos-based orthogonal tensor SVD for big data reduction in cloud environment," *IEEE TBD*, vol. 5, no. 3, pp. 355–367, September 2019.

[13] H. Hacigümüs, B. Iyer, S. Mehrotra, and C. Li, "Executing SQL over encrypted data in the database-service-provider model." in *Proc. of SIGMOD*, Madison, WI, USA, June 2002.

[14] R. Popa, C. Redfield, N. Zeldovich, and H. Balakrishnan, "CryptDB: Protecting confidentiality with encrypted query processing," in *Proc. of SOSP*, Cascais, Portugal, October 2011.

[15] S. De Capitani di Vimercati, R. Erbacher, S. Foresti, S. Jajodia, G. Livraga, and P. Samarati, "Encryption and fragmentation for data confidentiality in the cloud," in *FOSAD*, A. Aldini, J. Lopez, and F. Martinelli, Eds. Springer, 2014.

[16] M. Naveed, S. Kamara, and C. V. Wright, "Inference attacks on property-preserving encrypted databases," in *Proc. of ACM CCS*, Denver, CO, USA, October 2015.

[17] J. Duchi, M. Jordan, and M. Wainwright, "Local privacy and statistical minimax rates," in *Proc. of FOCS*, Berkeley, CA, USA, October 2013.

[18] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *IJUFKS*, vol. 20, no. 6, pp. 793–817, December 2012.

[19] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and verifiably encrypted signatures from bilinear maps," in *Proc. of Eurocrypt*, Warsaw, Poland, May 2003.

[20] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in *Proc. of ACM CCS*, Alexandria, VA, USA, October/November 2007.

[21] A. Juels and B. Kaliski, "PORs: Proofs of retrievability for large files," in *Proc. of ACM CCS*, Alexandria, VA, USA, October–November 2007.

[22] A. Juels and A. Oprea, "New approaches to security and availability for cloud data," *Communications of ACM*, vol. 56, no. 2, pp. 64–73, February 2013.

[23] G. Singh, S. Bharathi, A. Chervenak, E. Deelman, C. Kesselman, M. Manohar, S. Patil, and L. Pearlman, "A metadata catalog service for data intensive applications," in *Proc. of SC*, Phoenix, AZ, USA, November 2003.

[24] L. Ren, C. W. Fletcher, A. Kwon, E. Stefanov, E. Shi, M. van Dijk, and S. Devadas, "Constants count: Practical improvements to Oblivious RAM," in *Proc. of USENIX*, Washington, DC, USA, August 2015.

[25] E. Stefanov, M. van Dijk, E. Shi, C. W. Fletcher, L. Ren, X. Yu, and S. Devadas, "Path ORAM: An extremely simple Oblivious RAM protocol," in *Proc. of ACM CCS*, Berlin, Germany, November 2013.

[26] S. De Capitani di Vimercati, S. Foresti, S. Paraboschi, G. Pelosi, and P. Samarati, "Three-server swapping for access confidentiality," *IEEE TCC*, vol. 6, no. 2, pp. 492–505, April-June 2018.

[27] ——, "Shuffle index: Efficient and private access to outsourced data," *ACM TOS*, vol. 11, no. 4, pp. 19:1–19:55, October 2015.

[28] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati, "An authorization model for multi-provider queries," *Proc. of the VLDB Endowment*, vol. 11, no. 3, pp. 256–268, November 2017.

[29] C. Ardagna, S. De Capitani di Vimercati, S. Foresti, S. Paraboschi, and P. Samarati, "Minimising disclosure of client information in credential-based interactions," *IJIPSI*, vol. 1, no. 2/3, pp. 205–233, 2012.

[30] C. Ardagna, J. Camenisch, M. Kohlweiss, R. Leenes, G. Neven, B. Priem, P. Samarati, D. Sommer, and M. Verdicchio, "Exploiting cryptography for privacy-enhanced access control: A result of the PRIME project," *JCS*, vol. 18, no. 1, pp. 123–160, January 2010.

[31] C. Ardagna, M. Cremonini, S. De Capitani di Vimercati, and P. Samarati, "A privacy-aware access control system," *JCS*, vol. 16, no. 4, pp. 369–392, September 2008.

[32] E. Arfelt, D. Basin, and S. Debois, "Monitoring the GDPR," in *Proc. of ESORICS*, Luxembourg, September 2019.

[33] E. Birrell, A. Gjerdrum, R. van Renesse, H. Johansen, D. Johansen, and F. Schneider, "SGX enforcement of use-based privacy," in *Proc. of WPES*, Toronto, Canada, October 2018.

[34] P. Samarati and S. De Capitani di Vimercati, "Access control: Policies, models, and mechanisms," in *FOSAD*, ser. LNCS, R. Focardi and R. Gorrieri, Eds. Springer-Verlag, 2001.

[35] K. Frikken, M. Atallah, and J. Li, "Attribute-based access control with hidden policies and hidden credentials," *IEEE Transactions on Computers*, vol. 55, no. 10, pp. 1259–1270, October 2006.

[36] J.-W. Byun and N. Li, "Purpose based access control for privacy protection in relational database systems," *The VLDB Journal*, vol. 17, no. 4, pp. 603–619, July 2008.

[37] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati, "Encryption policies for regulating access to outsourced data," *ACM TODS*, vol. 35, no. 2, pp. 12:1–12:46, April 2010.

[38] M. Atallah, M. Blanton, N. Fazio, and K. Frikken, "Dynamic and efficient key management for access hierarchies," *ACM TISSEC*, vol. 12, no. 3, pp. 18:1–18:43, January 2009.

[39] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in *Proc. of PKC*, Taormina, Italy, March 2011.

[40] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: Pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132 – 15154, February 2018.

[41] H. S. Galal and A. M. Youssef, "Verifiable sealed-bid auction on the ethereum blockchain," in *Proc. of FC*, Curaçao, Netherlands, 2018.

[42] R. Cheng, F. Zhang, J. Kos, W. He, N. Hynes, N. Johnson, A. Juels, A. Miller, and D. Song, "Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contract execution," in *Proc. of IEEE EuroS&P*, Stockholm, Sweden, June 2018.

[43] S. Delgado-Segura, C. Pérez-Sola, G. Navarro-Arribas, and J. Herrera-Joancomartí, "A fair protocol for data trading based on bitcoin transactions," *FGCS*, 2017, pre-print.

[44] S. Dziembowski, L. Eckey, and S. Faust, "FairSwap: How to fairly exchange digital goods," in *Proc. of ACM CCS*, Toronto, Canada, October 2018.

[45] S. Rathore, Y. Pan, and J. Park, "BlockDeepNet: A blockchain-based secure deep learning for IoT network," *Sustainability*, vol. 11, no. 14, pp. 1–15, July 2019.